USER GUIDE

This document presents a user guide of CNAtra, a MATLAB-based tool for detecting large structural-variations and focal amplifications/deletions of low-coverage whole genome sequencing (WGS) data, especially for cancer cell lines. CNAtra is a software package developed using **MATLAB R2016b**.

CNAtra output at a glance

User can run CNAtra tool and get the CNV profile quickly using the default parameters if the BAM file is available. We provided the simulated data where we artificially incorporated LCVs and FAs in Chr3 of CHP-212 cell line data as an example for testing CNAtra installation. Input BAM is available as "**CHP212_chr3_Artificial.bam**" file under "**CNAtraInput**" folder. CNAtra outputs were provided in "**CNAtraResults**" folder (default output directory). After cloning CNAtra git repository, user can run the following commands using MATLAB to get the CNV profile.

```
>> CNAtraDirectory = './CNAtraTool';
>> addpath(CNAtraDirectory);
>> inputFile = './CNAtraInput/CHP212_chr3_Artificial.bam';
>> CNAtraObj = CNAtra(inputFile, CNAtraDirectory);
>> CNAtraObj.RDcalculator;
>> CNAtraObj.CNVcaller;
```

Quick start of CNAtra

In this Quick Start section, we briefly explain how you can install and use the CNAtra tool using MATLAB R2016b or any newer version. We will use IMR-32 cell line **(GEO accession number: GSE90683/GSM2664328)** as a test example in this document. Default annotation files such as GC-contents and mappability are given for read length of 100bps. For other read lengths, it is recommended to set their annotation files as explained below.

1. **Input preparation:** CNAtra accepts BAM/SAM files as input for RD calculation. It can handle both single-end and paired-end reads. Input BAM file is recommended to be sorted by position and indexed. Otherwise, MATLAB BioMap class may take longer time and space to sort it. You can use “samtools” for sorting and indexing the input BAM file.
   - *samtools sort input.bam > input.sorted.bam*
   - *samtools index input.sorted.bam*

2. **CNAtra setup:** download and extract CNAtra tool “CNAtraTool.zip” at any directory. Then, you can launch and add CNAtraTool directory to the top of the search path for the current MATLAB session.

   ```
   >> CNAtraDirectory = './CNAtraTool';
   >> addpath(CNAtraDirectory);
   ```

3. **Input load**: create a CNAtra object and load the input BAM/SAM file.

>> inputFile = './Data/IMR32.bam';
>> IMR32 = CNAtra(inputFile, CNAtraDirectory);
"Now, a CNAtra object (IMR32) is created with the default parameters."

4. **Read-depth Calculator**: compute the RD signal at 1Kb bin from the input sequence alignment data after removing low-quality reads. Then, the RD signal is normalized to remove the GC-biases, create a CNAtra object, and load the input BAM/SAM file. You can choose the GC-correction method (***gcCorrectionMethod)*** and set map-quality threshold (***MAPQ***) (*see CNAtra Parameters*). Finally, CNAtra default parameters are calculated based on data coverage.

>> IMR32.RDcalculator;
"Now, a CNAtra object is updated with coverage-based threshold parameters, RD signals, mappability tracks, and GC tracks."

>> IMR32

```
>> IMR32

IMR32 =

  CNAtra with properties:

                    bamFile: './Data/IMR32.bam'
              gcWindsMatFile: './CNAtraTool/referenceFiles/ChrisMillerGCs/gcWinds.hg19.readLength100.mat'
                  mapMatFile: './CNAtraTool/referenceFiles/AnshulMappability/mapTracks.hg19.101.mat'
                     gapFile: './CNAtraTool/referenceFiles/hg19.gaps.bed'
               blackListFile: './CNAtraTool/referenceFiles/wgEncodeDacMapabilityConsensusExcludable.bed'
             centromeresFile: './CNAtraTool/referenceFiles/h19centromeres.txt'
              telomeresFile: './CNAtraTool/referenceFiles/h19telomeres.txt'
            outputDirectory: './CNAtraResults'
            memoryFootPrint: 100000
                 MAPQ_Score: 1
          gcCorrectionMethod: 1
                ploidyLevel: 'free'
              minimumIBsize: 1000
                 resolution: 40
     amplificationThreshold: 0.770000000000000
          deletionThreshold: 0.747000000000000
          minAlterationRank: 2
    minMappabilityThreshold: 0.500000000000000
    maximumFalseBinsAllowed: 0.500000000000000
               numberOfReads: 42090785
                dataCoverage: 1.400106586038952
                  readLength: 101
                    readType: 'Single-end'
             chrRawDictionary: [23×1 containers.Map]
                chrDictionary: [23×1 containers.Map]
               chrFDictionary: [23×1 containers.Map]
                    chrFIndex: [23×1 containers.Map]
                    chrNames: [23×1 containers.Map]
                  chrLengths: [23×1 containers.Map]
       chrMappabilityTracks: [23×1 containers.Map]
                 chrGCTracks: [23×1 containers.Map]
     chrCentroTeloBoundaries: [23×1 containers.Map]
    chrBlackListedBoundaries: [23×1 containers.Map]
             chrGapBoundaries: [23×1 containers.Map]
                 CNReference: []
              segmentsInfoDic: []
               regionsInfoDic: []
```

5. **CNV Caller**: identify the iso-copy number blocks (IBs) and the focal amplifications/deletions (FAs). Output files are generated including their copy numbers, significant classes, and other attributes as the distance to the centromere, telomeres, and edges of iso-copy numeric blocks (*see CNAtra Output*). In addition, CNAtra generates statistics of alteration types and classes. Also, it saves BED format files that can be uploaded into the UCSC Genome Browser from individual altered region visualization. Before executing the CNV Caller module, user can change the CNAtra pipeline parameters (see CNAtra Parameters) such as the estimated coverage-based thresholds, resolution, and the filtering parameters **(*amplificationThreshold,* ploidyLevel, *…)***.

>> IMR32.CNVcaller;

"Now, a CNAtra object is updated with the both IBs and focal amplifications/deletions."

>> IMR32

```
>> IMR32

IMR32 =

  CNAtra with properties:

                    bamFile: './Data/IMR32.bam'
               gcWindsMatFile: './CNAtraTool/referenceFiles/ChrisMillerGCs/gcWinds.hg19.readLength100.mat'
                   mapMatFile: './CNAtraTool/referenceFiles/AnshulMappability/mapTracks.hg19.101.mat'
                      gapFile: './CNAtraTool/referenceFiles/hg19.gaps.bed'
                blackListFile: './CNAtraTool/referenceFiles/wgEncodeDacMapabilityConsensusExcludable.bed'
              centromeresFile: './CNAtraTool/referenceFiles/h19centromeres.txt'
                telomeresFile: './CNAtraTool/referenceFiles/h19telomeres.txt'
              outputDirectory: './CNAtraResults'
              memoryFootPrint: 100000
                   MAPQ_Score: 1
           gcCorrectionMethod: 1
                  ploidyLevel: 'free'
               minimumIBsize: 1000
                  resolution: 40
       amplificationThreshold: 0.770000000000000
            deletionThreshold: 0.747000000000000
           minAlterationRank: 2
      minMappabilityThreshold: 0.500000000000000
      maximumFalseBinsAllowed: 0.500000000000000
                numberOfReads: 42090785
                 dataCoverage: 1.400106586038952
                   readLength: 101
                     readType: 'Single-end'
             chrRawDictionary: [23×1 containers.Map]
                chrDictionary: [23×1 containers.Map]
               chrFDictionary: [23×1 containers.Map]
                    chrFIndex: [23×1 containers.Map]
                    chrNames: [23×1 containers.Map]
                  chrLengths: [23×1 containers.Map]
         chrMappabilityTracks: [23×1 containers.Map]
                  chrGCTracks: [23×1 containers.Map]
      chrCentroTeloBoundaries: [23×1 containers.Map]
     chrBlackListedBoundaries: [23×1 containers.Map]
             chrGapBoundaries: [23×1 containers.Map]
                  CNReference: 1.449388925163155e+03
              segmentsInfoDic: [23×1 containers.Map]
               regionsInfoDic: [23×1 containers.Map]
```

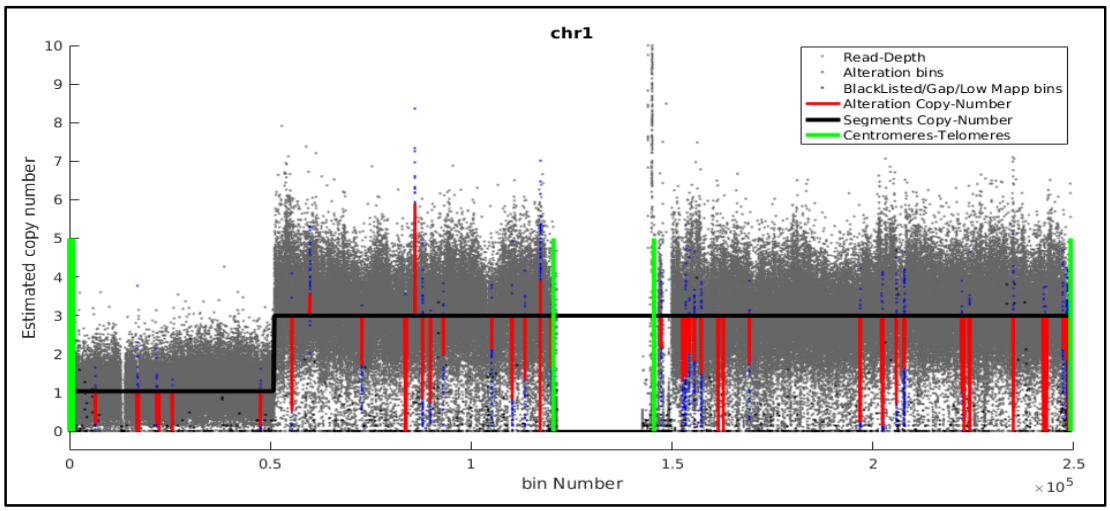
6. **Visualization**: plot (save) figures of the copy number profile of a genomic region. This region can be a chromosome, iso-copy numeric block (IB), focal amplification/deletion, or any specific region defined by the genome coordinates. This can be used for visual assessment of IBs and focal amplifications/deletions. In addition, the user can plot the ploidy test of the cell line.

Method	Function
CNAPlot(*option, chrNo****)*** ***CNAPlot(****option, chrNo, IB****)*** ***CNAPlot(****option, chrNo, startPos, stopPos****)***	plot the copy number profile of a chromosome, IB, or genome region. - *option*: 'plot' for plotting a figure of the specified genome region, 'save' for saving this figure. - *chrNo*: the chromosome number from 1 to 23. 23 is chromosome X. - IB: the order of the IB per chromosome to be plotted. - *startPos*: the start position of the selected genome region to be plotted. - *stopPos*: the stop position of the selected genome region to be plotted.
plotGenome(*binSize****)***	plot the genome-wide RD signal with the estimated copy number and the IBs. - *binSize*: the bin size for plotting the RD signal (minimum 1000 bp).
CNARegionPlot*(option, chrNo, regionNo)*	plot the copy number profile a genomic region centralized around focal amplification/deletion using its order in the chromosome. - *option*: 'plot' for plotting a figure of the specified genome region, 'save' for saving this figure. - *chrNo*: the chromosome number from 1 to 23. 23 is chromosome X. - *regionNo*: the order of the focal amplification/deletion region per chromosome to be plotted.
CNVsTrackPlot(*option****,*** *chrNo****)***	plot the CNV track of a chromosome by merging both the IBs and FAs. - *option*: 'plot' for plotting a figure of the specified genome region, 'save' for saving this figure. - *chrNo*: the chromosome number from 1 to 23. 23 is chromosome X.
CNAEstimator(*chrNo, startPos, stopPos****)***	return the estimated copy number of a genome region. - *chrNo*: the chromosome number from 1 to 23. 23 is chromosome X. - IB: the order of the IB per chromosome to be plotted. - *startPos*: the start position of the selected region. - *stopPos*: the stop position of the selected region.

ploidyTest	plot the histogram of IBs' copy numbers versus the multimodal distribution to validate the ploidyLevel assumption.

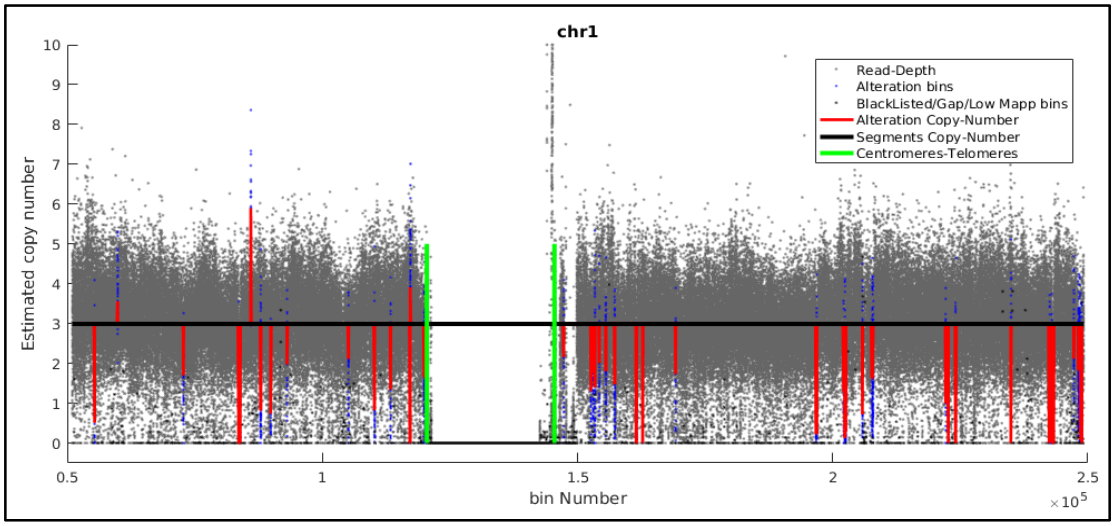
Ex1: plot the copy number profile of chromosome 1 of IMR32

>> IMR32.CNAPlot('plot',1);



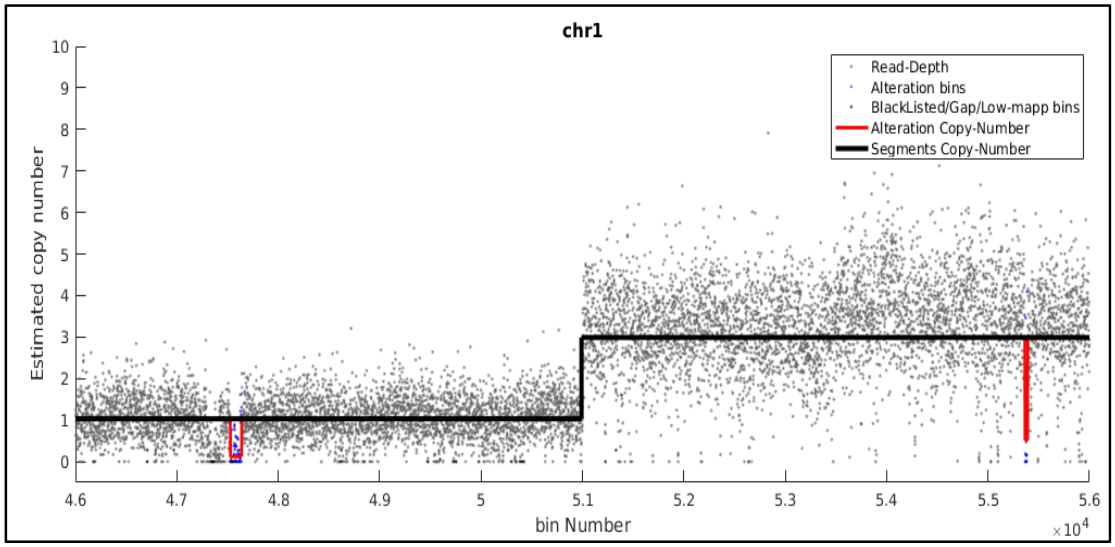
Ex2: plot the copy number profile of IB 2 of chromosome 1 of IMR32

>> IMR32.CNAPlot('plot',1, 2);



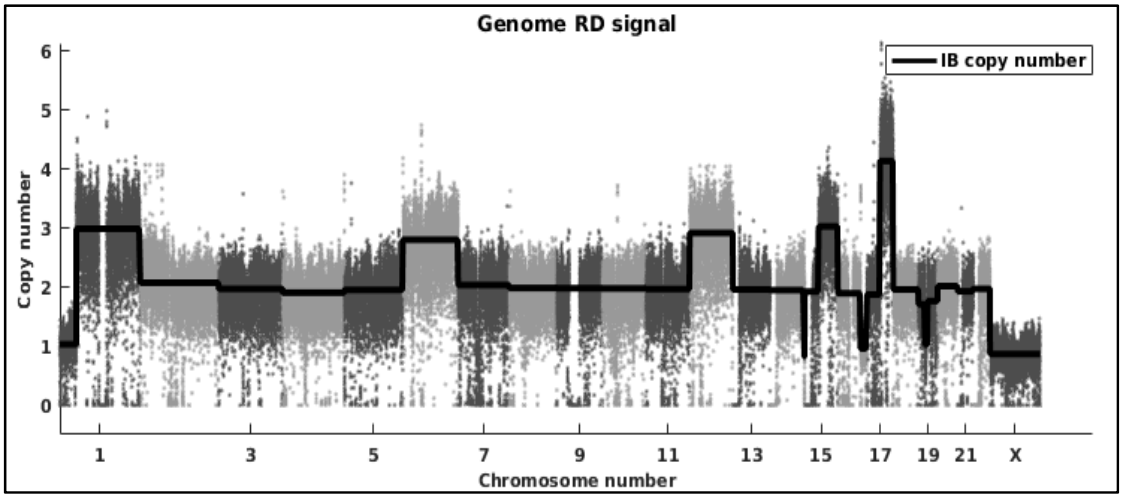
Ex3: plot the copy number profile of [46Mb: 56Mb] region of chromosome 1 of IMR32

>> IMR32.CNAPlot('plot',1, 46000000, 56000000);



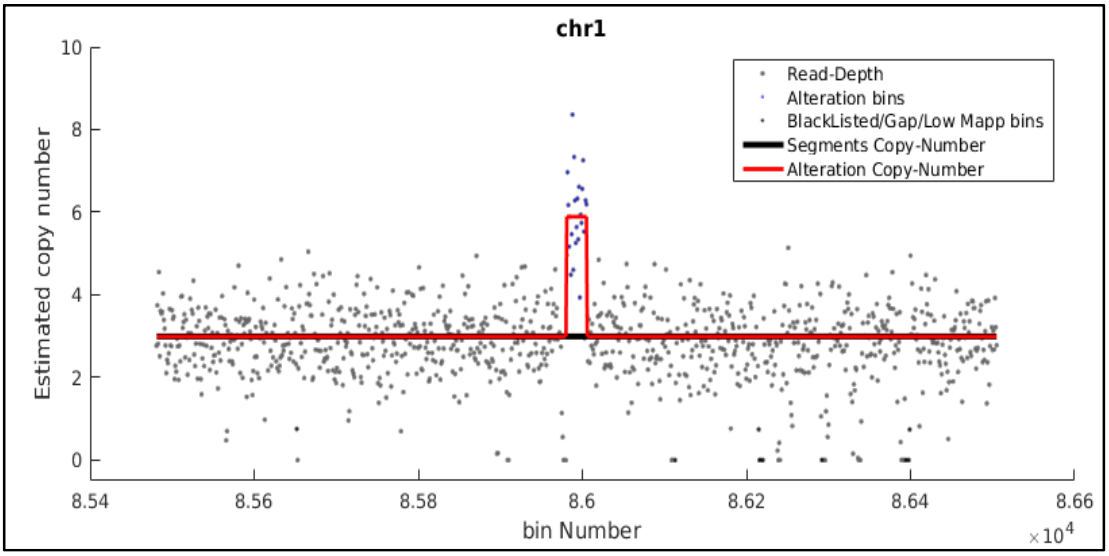
Ex4: plot the genome RD signal using a bin size of 20Kb

>> IMR32.plotGenome(20000);



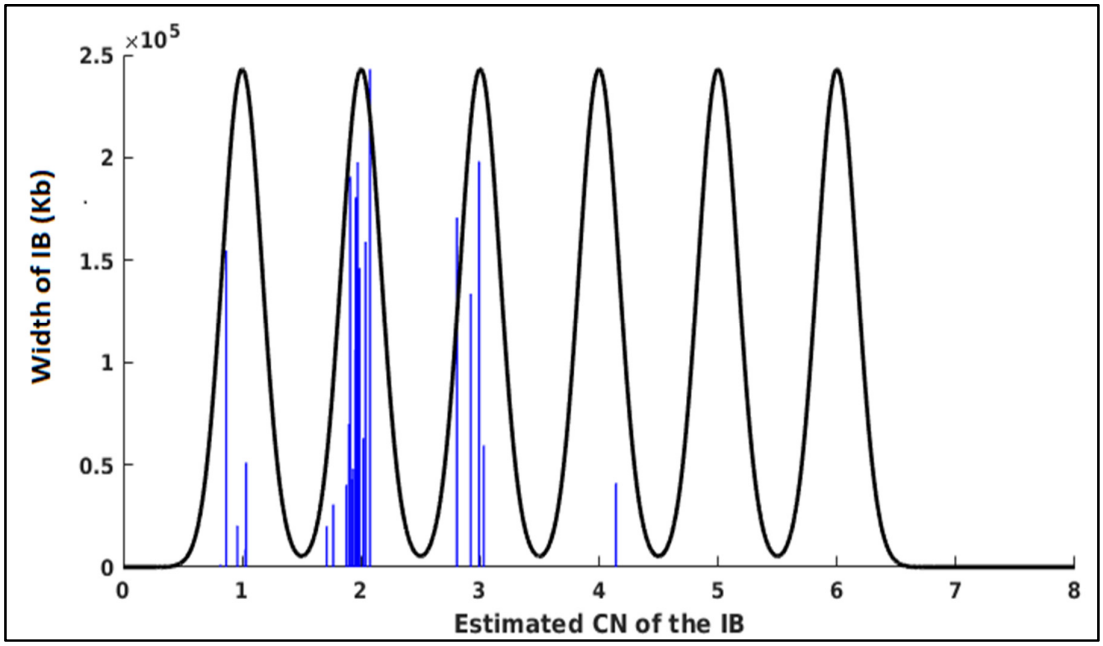
Ex5: plot the focal amplification/deletion region (number = 2 in the chromosome order) of chromosome 1 of IMR32

>> IMR32.CNARegionPlot('plot',1, 2);



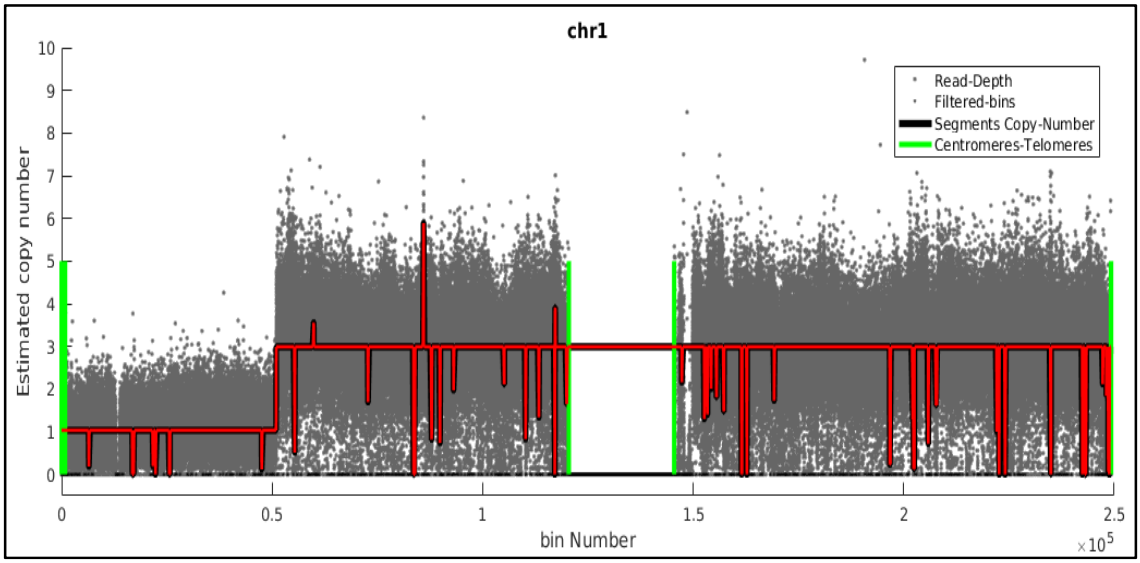
Ex6: plot the histogram of the copy numbers of IBs versus the multimodal distribution.

>> IMR32.ploidyTest;



Ex7: plot the CNV track of chromosome 1 of IMR32

>> IMR32.CNVsTrackPlot('plot', 1); %plot the CNV track of chromosome 1.

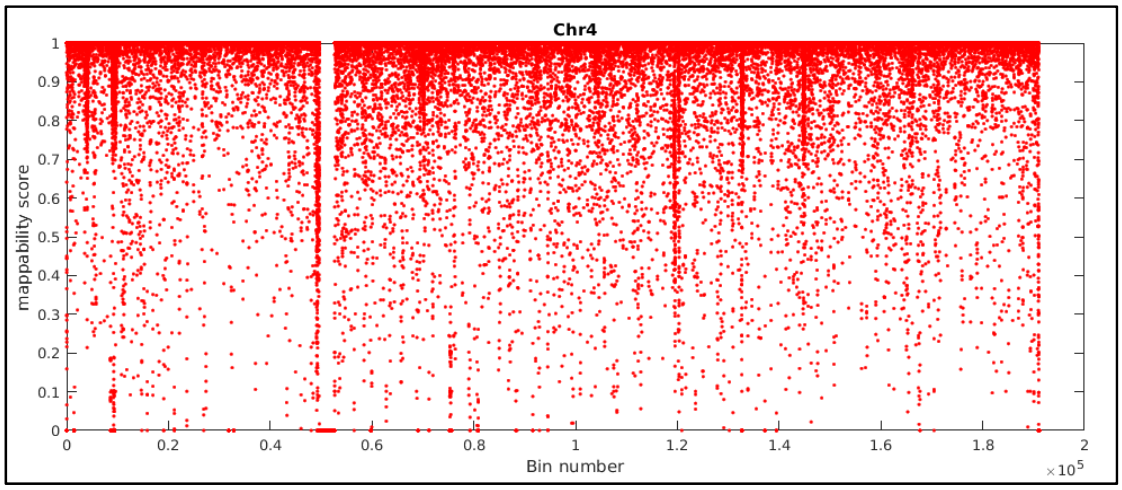


"Also, user can also manually plot the RD-signal/mappability track/GC track of any chromosome by accessing their dictionaries."

Dictionary	Description
chrRawDictionary	- dictionary of raw RD signal per chromosome before correction and filtering. For example, **chrRawDictionary(1)** is the raw RD signal of the chromosome 1. Chromosomes can be any value from 1 to 23 (23: chromosome X).
chrDictionary	- dictionary of RD signal per chromosome after GC and mappability correction.
chrFDictionary	- dictionary of RD signal per chromosome after filtering the low-mappability, black-listed, and gap regions.
chrGCTracks	- dictionary of GC-track per chromosome loaded from Chris Miller's GC tracks.
chrInputGCContents	- dictionary of GC-track per chromosome computed from the input reads.
chrMappabilityTracks	- dictionary of mappability-track per chromosome loaded from Anshul Kundaje's mappability tracks.

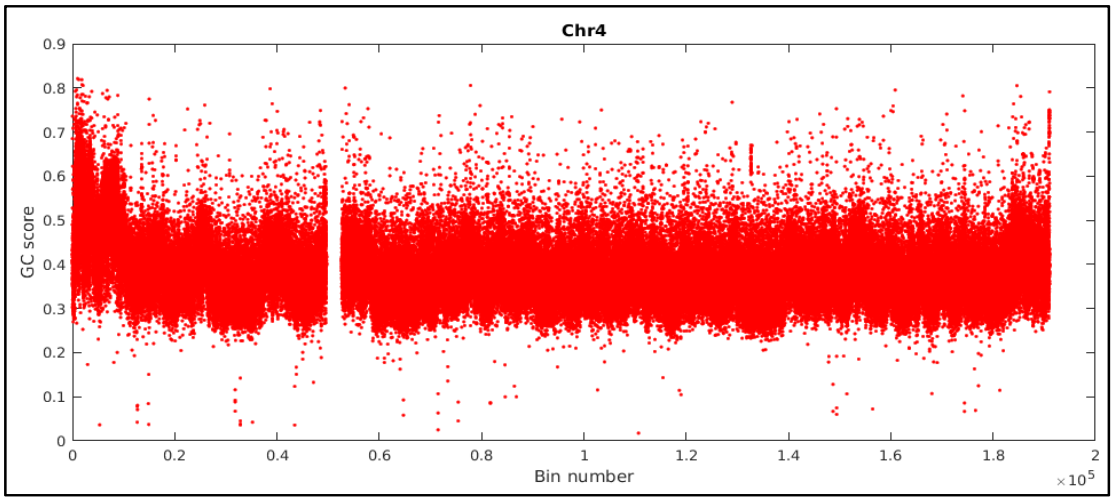
Ex8: plot the mappability-tracks of chromosome 4

```
>> figure;
>> chrNumber = 4;
>> plot(IMR32.chrMappabilityTracks(chrNumber),'r.');
>> xlabel('Bin number');
>> ylabel('Base count');
>> title('Chr4 mappability scores');
```



Ex9: plot the GC-tracks of chromosome 4

>> figure;
>> chrNumber = 4;
>> plot(IMR32.chrGCTracks(chrNumber),'r.');
>> xlabel('Bin number');
>> ylabel('Base count');
>> title('Chr4 GC scores');



7. **Object Load/Store**: user can store the **CNAtra** object including all attributes and parameters (RD signal, IBs, focal alterations, …) using the MATLAB ***save*** method. Then, user can load it for future analysis without the need to run the whole pipeline again.

   >> save('IMR32_CNAtraObject.mat', 'IMR32');

   "Now, a CNAtra object can be loaded anytime directly by the following commands."

   >> CNAtraDirectory = './CNAtraTool';

   >> addpath(CNAtraDirectory);

   >> load('IMR32_CNAtraObject.mat', 'IMR32');

CNAtra Parameters

User can set CNAtra default parameters before executing the CNAtra methods such as ***RDcalculator*** and ***CNVcaller***. For example, a user can set MAPQ_Score, gcCorrectionMethod before running the ***RDcalculator***. Also, for different read length, you can change default annotation files such as gcWindsMatFile and mapMatFile since the default files are given for read length of 100bps. So, read depth calculation is carried based on these parameters and annotation files. Similar, user can set the estimated parameters such as *minAlterationRank, amplificationThreshold, ploidyLevel, and resolution.* It is recommended to set minAlterationRank to 3 for complex data profile such as cancer cell lines. User can change the ploidyLevel based on the expected ploidy level of the cell line "haploid, diploid, triploid, tetraploid."

Parameter	Description
User should set their values before running the 'RDcalculator' module if needed.	
MAPQ_Score	The threshold for filtering the reads from the BAM file. **The default value is 1** assuming Bowtie2 mapping. However, the user can change it based on his short sequence mapper.
memoryFootPrint	Number of reads per iterations for GC-calculations. User can decrease it if there is a memory error during executing the **RDcalculator** module (**default = 100000**).
gcCorrectionMethod	The method for correcting the read-depth value per bin. **0**: No GC-correction is applied, **1**: Genome-wise GC-correction using pre-calculated Chris Miller's GC tracks (**default**), **2**: Genome-wise GC-correction using the GC tracks that are calculated from the input data.
minMappabilityThreshold	The minimum mappability-score of a bin to be kept for CNAtra analysis. If the user set it to 1, **minMappabilityThreshold** is recomputed as 10thPercentile of the mappability scores based on read length (**default = 0.5**).
User should set their values before running the 'RDcalculator' module if needed.	
gcWindsMatFile	File that contains the precalculated GC scores per bin. A default hg19 file "**CNAtraDirectory+/referenceFiles/ChrisMillerGCs/gcWinds.readLength100.mat**"
mapMatFile	File that includes mappability scores of the genome (at 1Kb bins) which can be used to filter highly repeated or unmappable regions. A default hg19-based file "**CNAtraDirectory+/referenceFiles/AnshulMappability/mapTracks.hg19.101.mat**" is included by default if user doesn't specify one. This file can be used for single-end reads with read length >= 100bps. For other read lengths, it is better to load other files.
blackListFile	File that includes black-listed regions. A default hg19 file "**CNAtraDirectory+/referenceFiles/wgEncodeDacMapabilityConsensusExcludable**" is included if user doesn't specify one.
centromeresFile	File that includes locations of centromeres. A default hg19 file "**CNAtraDirectory+/referenceFiles/h19centromeres.txt**" is included.
telomeresFile	File that includes locations of telomeres. A default hg19 file "**CNAtraDirectory+/referenceFiles/h19telomeres.txt**" is included.
gapFile	File that includes unmappable gab hg19 regions. A default hg19-based file "**CNAClassDirectory+/referenceFiles/hg19.gaps.bed**" is included.

User should set their values before running the 'CNVcaller' module and after executing the 'RDcalculator' module if needed.	
amplificationThreshold	Threshold for identifying the **Class2** focal amplification.
deletionThreshold	Threshold for identifying the **Class2** focal deletions.
resolution	Minimum width of focal amplification/deletion to be kept.
ploidyLevel	Expected number of chromosome sets (default = 'free'), user can set it based on input-cell ploidy {'diploid', 'triploid', 'tetraploid'}.
minAlterationRank	Minimum rank of focal amplification/deletion to be considered as significant region (**default value = 2**). **2**: {'**Class1**', '**Class2**'}, **5**: {'**Class2**'} only.
MaximumFalseBinsAllowed	Maximum ratio of gab/low-mappability/black-listed bins per alteration region to be kept as significant one (default = 0.5).

Ex1: set GC-Correction method to 2 to calculate the GC-tracks from the input reads and re-rum RDcalculator method

>> IMR32.gcCorrectionMethod = 2;
>> IMR32.RDcalculator;

Ex2: set minimum alteration rank to 5 to filter out class1 FAs and re-run CNVcaller method

>> IMR32.minAlterationRank = 5;
>> IMR32.CNVcaller;

Ex3: set ploidy level to "triploid" to be used to compute the copy number reference and re-run CNVcaller method

>> IMR32.ploidyLevel = 'diploid';
>> IMR32.CNVcaller;

CNAtra Supplementary Files

We use 'gcWindsMatFile' and 'mapMapFile' input MATLAB files that contain the GC-tracks and mappability tracks for 1Kb bins. Our default files are set based on 100bp read length. However, we also provided the GC-tracks and mappability tracks for other common read lengths (27, 36, 50, 76) in " **CNAtraDirectory+/referenceFiles/ChrisMillerGCs"** and **"CNAtraDirectory+/referenceFiles/AnshulMappability"** folders respectively. It is recommended to set "**gcWindsMatFile"** and "**mapMatFile"** based on the read length for accurate filtering and normalization. For other read lengths, we also provided a MATLAB scripts in these folders for creating the similar MATLAB input files.

CNAtra Output

CNAtra generates many output files providing the detailed characterization of the copy number profile of the input cell line. By default, all results and figures are saved in the ***outputDirectory*** folder (**default = './CNAtraResults'**); user can change it by setting the **outputDirectory** attribute of the **CNAtra object**.

a. CNAtra focal amplification & deletions: we provide a **narrowPeak BED format file** of the focal amplifications /deletions of the complete genome for UCSC Genome Browser:

Column number	Description
1	Chromosome number
2	The starting position of the focal amplification/deletion
3	The ending position of the focal amplification/deletion
4	The class of the focal region {'Class2, 'Class1'}
5	Grey-scale score of the copy number of the focal region.
6	The type of the focal region {'+': amplification, '-': deletion}
7	The copy number of the focal amplification/deletion.
8	The ID (order) of the focal amplification/deletion per chromosome.

Ex1: IMR32 focal amplifications/deletions

```
track type=narrowPeak visibility=3 description="CNAtra focal amplifications/deletions
chr1    6481000 6528999 Class2  2.170466e+01    -   0.1852  1   -1  -1
chr1    16819000    17227999    Class2  1   -   0.0000  2   -1  -1
chr1    21708000    21819999    Class2  2.283895e+01    -   0.1919  3   -1  -1
chr1    22294000    22347999    Class2  1   -   0.0000  4   -1  -1
chr1    25586000    25664999    Class2  1   -   0.0000  5   -1  -1
chr1    25695000    25754999    Class2  1   -   0.0000  6   -1  -1
chr1    47527000    47638999    Class2  1.441006e+01    -   0.1205  7   -1  -1
chr1    55363000    55390999    Class2  6.123140e+01    -   0.5116  8   -1  -1
```

b. CNAtra IBs: we provide a **narrowPeak BED format file** of the iso-copy numeric blocks of the complete genome for UCSC Genome Browser:

Column number	Description
1	Chromosome number
2	The starting position of the IB.
3	The ending position of the IB.
4	Tag of the IB (ex: IB1, IB2)
5	Grey-scale score of the copy number of the IB.
6	.
7	The copy number of the IB.
8	The ID (order) of the IB per chromosome.

Ex2: IMR32 IBs

```
track type=narrowPeak visibility=3 description="CNAtra IBs"
chr1    0   50989999    IB1 103 .   1.0330  1   -1  -1
chr1    50990000    249196999   IB2 299 .   2.9911  2   -1  -1
chr2    0   242995999   IB3 207 .   2.0748  1   -1  -1
chr3    0   197817999   IB4 197 .   1.9711  1   -1  -1
chr4    0   190791999   IB5 190 .   1.9076  1   -1  -1
chr5    0   180679999   IB6 195 .   1.9540  1   -1  -1
```

c. CNAtra cell line statistics: we provide a text file providing the complete statistics of focal amplifications/deletions, their classes, and IBs for each chromosome:

Column number	Description
1	Chromosome number
2	The total number of IBs per chromosome.
3	The total number of focal amplifications/deletions per chromosome.
4	The number of focal amplifications per chromosome.
5	The number of focal deletions per chromosome.
6	The number of Class2 focal amplifications/deletions per chromosome.
7	The number of Class1 focal amplifications/deletions per chromosome.

Ex3: IMR32 statistics

Chromosome	#IBs		#CNVs		#Amplifications	#Deletions	Class2	Class1
chr1	2	48	3	45	47	1		
chr2	1	31	7	24	29	2		
chr3	1	29	3	26	26	3		
chr4	1	32	6	26	24	8		
chr5	1	27	2	25	24	3		

d. Individual chromosome results: we provide a detailed copy number profile (file for each chromosome) containing its focal amplifications/deletions, IBs, and other attributes that can be used for evaluation such as their types, significant classes, distances to centromere and telomeres and edges of iso-copy numeric blocks:

Ex4: IMR32 chromosome 1 copy number profile

Segment-Number	Start-Bin(Kb)	Stop-Bin(Kb)	Segment-Width(Kb)	Copy-Number
1	1	50990	50990	1.033003

ID	Start(bp)	Stop(bp)	Width(Kb)	Copy-Number	Class	SegmentEdge Distance(Kb)	Centro/Telo Distance(Kb)
1	6481000	6528999	48	0.18523	Class2	6481	5635
2	16819000	17227999	409	0.00000	Class2	16819	15973
3	21708000	21819999	112	0.19195	Class2	21708	20862
4	22294000	22347999	54	0.00000	Class2	22294	21448
5	25586000	25664999	79	0.00000	Class2	25325	24740
6	25695000	25754999	60	0.00000	Class2	25235	24849
7	47527000	47638999	112	0.12050	Class2	3351	46681